The use of relational databases for electronic and conventional scientific publishing

Joost G. Kircz
North-Holland Physics Publishing Division, Elsevier Science Publishers B.V., P.O. Box 103, 1000 AC Amsterdam, The Netherlands

Jan Bleeker
Automation Department, Elsevier Science Publishers B.V. Amsterdam, The Netherlands

Received 11 July 1986

In an attempt to integrate traditional scientific publishing and new electronic possibilities for storage, retrieval and dissemination, a model is proposed. Based on a General Mark-up Language approach the information content of an article is analyzed and various categories of different types of information are defined. Using a relational database as a starting point for all operations, different products can be identified by packaging the different information entities in a multitude of ways.

Traditional publishing results by aggregating the database fields in a typeset file; secondary information (bibliographic and citation indexing) can be extracted automatically whilst all kinds of publisher's statistics ranging from geographical breakdowns or author's lists to subject indexes, are spin-offs as well. A new feature is a current awareness service which enables the user to search for literature prior to publication. Linked with the storage of the complete manuscript, including figures and tables on, e.g. an optical disc, this enables the possibility of linking bibliographic searching to full text retrieval. Recommendations for an author's environment to be linked to such an integrated publisher's model are discussed.

1. Introduction

In scientific publishing, the continued increase in the number of publications, especially research articles, is generating a number of new and difficult problems. In scientific research there are recognized changes in the publication patterns in the postwar period compared with the entire history of scientific publishing before the second world war. The acceleration of research has made it very difficult for individuals to be able to keep abreast of an entire field. Increasing specialization develops new sub-disciplines that, in turn, develop their own specialized books and journals.

While, on the one hand the publications become more and more specialized, on the other the judging of the quality of research by the outside world (e.g. funding agencies) is to some extent influenced by the number of publications researchers produce. Therefore, in the world of scientific communications, we see the following:

(a) articles tend to become more and more specialized;
(b) the number of articles tends to grow;
(c) owing to the growing competition for funds and the general acceleration of research, the ability to digest all published research becomes more difficult;
(d) authors have a strong desire for faster publication in order to lay claim to their results as early as possible.

Given the abovementioned problems and given the permanent need for researchers themselves to have up-to-date information concerning the ongoing research of other groups during the course of their research programme, it is no exaggeration to claim that there is a serious information retrieval and dissemination problem in scientific research. The time lag between research and publication, as well as between publication and reading, leads to researchers receiving hot news via private communication, preprints and discussions at conferences.

Publication serves to a large extent the function of the final research report and proof to the outside world that, indeed, the research has been performed and has produced results. For the information industry and specifically scientific publishers, the occurrence of more publications, more difficult retrieval, and greater specialization of the different publications pose serious questions about the way they organize their activities. As information packagers, scientific publishers have a series of functions ranging from the acquisition, organization of editing, and refereeing, to the actual production in print and the distribution of the

North-Holland
Journal of Information Science 13 (1987) 75–89

research results. In order to cope with the flood of articles, the publishing industry has to review its activities accordingly, and see how the good aspects of the old-fashioned publications and the new possibilities provided by the electronic revolution can be merged [1]. It has to be realized that this is not only a problem of the type of information carrier. The capacity of a magnetic tape or a compact disc (CD-ROM) is much larger than a shelf of journal issues, but for the actual user this does not make much of a difference since the information still has to be selected and digested. The question is how can scientific publishers channel the information stream in such a way that retrieval and dissemination is guaranteed over an extensive period of time, and in such a way that this retrieval and dissemination is flexible (in the sense that future needs, as far as can be forecast, can be satisfied, as well as actual need and usage).

2. Standardization and scientific articles

Developing a storage and retrieval system for all possibilities of scientific information is a tremendous task. Therefore, we will try to take a pragmatic approach and deal only with the problems we think can be solved conceptually at the moment. For this reason, we will restrict ourselves to an analysis of scientific journal publishing. In research articles published in journals there is a clear historical tradition of presentation. Every scientific article has more or less the same structure. A functional division can be distinguished between different definable entities of information.

In general, every article carries the names of the authors and a title as well as the body of the article itself and bibliographic references. Most articles have an abstract or summary and have the authors' affiliations with, or without, the addresses. Some articles have tables and/or figures, other articles have, e.g. mathematical or chemical formulae. Given such a consistency it is possible to analyze the structure of a scientific journal paper and to work out a way of sorting and storing the information to enable the maximum variety of possible final presentations. We will try to analyze below the structure of scientific journal articles. We will define information entities, and this analysis will lead us to a conceptual approach for all scientific article publishing activities.

2.1. Information structure and technical structure

Besides the structure of scientific information provided in an article we also have to deal with its technical structure. By technical structure we mean: typesetting instructions, if the articles are printed; file and database codes, if articles are on electronic storage devices and eventually distributed electronically, etc. To process all these differently identifiable entities, we have to use a generalized approach to all the information contained in a journal article. For that purpose we have to consider not only typesetting codes for symbols, non-standard alphabets, etc., but, first and foremost, the development of a so-called Standard Generalized Mark-up Language (SGML) [2]. A Standard Generalized Mark-up Language is a method of describing documents in such a way that, in transferring them from one place to another, the information identification remains obvious. For example, if we transfer a file from one computer system to another, the tag 'title' will always indicate and be understood as a sign that the subsequent field will contain something which is considered to be a title. The same holds true in a traditionally printed index, where the alphabetically organized surnames of authors indicate that the field 'surname' has a well defined meaning and refers to articles published by an author who is identified by his/her surname. (A short explanation of the Standard Generalized Mark-up Language is given in Box 1.) It is important to note that the technical structure is, to a large extent, defined by the various types of information supplied by the article, e.g. the author's name is printed separately from the text body.

The generalized approach of analyzing articles into their smallest independent identifiable information parts makes it possible to package the information in the maximal number of ways in order to deal with a large variety of possible uses of the information (reading a whole article, checking the references, finding the street number of an institute, etc.). This method of element definition and the need for retrieval and selection of (parts of) text can be conveniently handled using the concept of relational database techniques in combination with fourth-generation software tools.

Box 1. Standard Generalized Mark-up Language

SGML stands for Standard Generalized Mark-up Language. SGML is the ISO standard description of text so that it can be stored in electronic media independently of the hardware or software being used for data capture (e.g. the wordprocessing package), and in such a way that it is also independent of the form in which it will ultimately be presented (by means of a typesetting system, via a database host, or via a microcomputer equipped with a compact disc reader, etc.). This sounds rather complicated, but it is no more than a set of rules for the description of texts, which in natural languages is called grammar.

SGML has two basic principles:

(1) The descriptors of texts (called SGML tags) must be based on content and not on form (the way it is presented), for that can be manifold. For example: a title must be named TITLE and not "bold, 10 points, etc.". Text parts that need emphasis, are called "emphasized" and not "italics". "Bold" and "italics" have to do with presentation, e.g. on paper, while in the case of display on a CRT it may be in 'reversed video' or 'underscored' instead of the way in which it is typeset.

(2) The SGML tags used for description of texts must be defined in a document description. This is based on the principle that texts are structured, i.e. independent of its purpose. One can describe the elements of which a text consists and in which order they have to be in the text, whether they are optional or obligatory and/or repetitive. This is especially true for scientific texts. Based on this document description, a text can be 'parsed', which means checked for the presence or absence of 'SGML tagged' text parts, as well as their order, etc. At the same time, a document description is used as a basis for the instructions being given for final presentation, e.g. to a typesetter. This is called a document presentation description, e.g. the contents after the SGML tag TITLE must be presented in "bold, 10 points, etc.".

Based on the SGML 'grammar', an SGML-application has been developed by Aspen Systems on behalf of the Association of American Publishers (AAP). AAP-SGML is the definition of a set of SGML tags and a set of document descriptions for text that will be published, e.g. as a book, an article in a journal, or a monograph [12].

This SGML application is in its final stage of development. The standard is being accepted by a majority of European publishers and therefore has a good chance of survival. Because addition of codes by the author will imply a lot of extra work, major manufacturers of word-processing packages will develop additional software tools that will assist authors with creation of SGML-coded texts.

In traditional publishing, databases are considered as mere spin-offs of the publishing activity and are created by using the published material, or the material in the last stage prior to final typesetting [3]. In our approach, we will start by taking the relational database as the pivot of all operations from the beginning of all activities, while all other endeavours will be considered as an application of the database. Of course, such a publishing *volte face* is not an obvious thing, as one application, i.e. typesetting and subsequent printing and distributing by normal mail of scientific information, will still make up 99% of the whole commercial activity. Nevertheless, we think that, conceptually, printed material has to be treated as an application just like database searching or having information available on other carriers like microfiche or compact disc.

In the analysis of different identifiable fields within a relational database, we do not want to deal with the maximum number of fields by dividing the articles into their smallest divisions. Such an atomistic approach will not bring any conceptual insight for the actual user. Most of the time, a person's goal is looking for relevant literature to scan or read completely. Therefore, our analysis will not identify all possible bits and pieces. It will attempt to identify the relevant parts of information in such a way that the total number of possible, searchable entities is within limits, structurally clear, and the flexibility of using the information is optimal. Hence, we will start our survey by analyzing the different information parts of the publishing process.

3. Information assembly, transfer and distribution

3.1. Author's information

(a) The information delivered by authors in their manuscripts can be analyzed in different ways. Our approach is to categorize this information in defined sets which play a distinct and different role in the course of information processing. A first category, which we will call

'first-order information' (Set {A}), is the information related to the *identification* of the article. This information consists of different sub-categories: author's name, which can again be split up into surname, given name, etc.; the author's address, which can be split into institute name, the town, the postal code, etc.; the title of the article; and, as a fourth part, the abstract of the article.

One can, of course, argue about the status of different sub-categories, but from a user's point of view, for all important aspects in retrieving scientific information, these four sub-categories are of primary interest. One cannot usually find a conventional article without a title or author's name; one cannot contact colleagues without knowing their address; and one usually makes a decision to read an entire article with some knowledge of the summary or abstract. In List 1 we have listed this 'first-order information' in more detail. This list (and all following lists) is not exhaustive, but our aim is only to have a clear overall picture.

List 1. First-order information
Set {A}. Identification
A1. Author name
  A1.1. Surname
  A1.2. First name(s) or initials
  A1.3. Pre or post particles (Sir, Jr., etc.)
A2. Author's address
  A2.1. Name of institute
    A2.1.1. University/corporation
    A2.1.2. Faculty/department
  A2.2. Street + number
  A2.3. Town
  A2.4. Postal code
  A2.5. Country
  A2.6. Telex
  A2.7. Electronic mailbox
  A2.8. Phone
  A2.9. Fax
A3. Title
  A3.1. 1st order subtitle
  A3.2. 2nd order subtitle
A4. Abstract
  A4.1. Elements of Set {B}

Note that in one article A1 and A2 might occur many times if more authors are involved.

(b) After the identification information, we have the information on the *text structure*—we call this 'second-order information' (Set {B}). This information on the text structure goes deeper into the details of the presentation of the article. Therefore, we can distinguish sub-categories of, e.g. the main text, which is a set of headings, paragraphs and related text. This related text can again be split up into information which is part of sections of the main text and part which is not. The first group contains things like quotations, indicators (so-called pointers) to references and footnotes, chemical formulae, mathematical formulae, tables (to be split up into table captions and the table itself), etc. The set of related text not belonging to sections consists of the footnotes themselves, the references, etc. On the last point, the references are again a world on their own, as they contain information on other articles, which can be partly described with the sub-categories of Set {A}, e.g. name and title. In List 2, a more elaborate division of 'second-order information' is given.

List 2. Second-order information
Set {B}. Text structure
B1. Main text
  B1.1. Sections
    B1.1.1. Section headings
    B1.1.2. Section body
      B1.1.2.1. Series of paragraphs
B2. Related text (part of sections)
  B2.1. Highlighted phrases
  B2.2. Quotations
  B2.3. Pointers
    B2.3.1. to footnotes
    B2.3.2. to references
    B2.3.3. to digitized figures
  B2.4. Lists
    B2.4.1. List header
    B2.4.2. List body
  B2.5. Tables
    B2.5.1. Captions
    B2.5.2. Contents
  B2.6. Boxes
    B2.6.1. Box header
    B2.6.2. Box body
  B2.7. Formulae
    B2.7.1. Chemical
    B2.7.2. Mathematical
B3. Related text (*not* part of sections)
  B3.1. Footnotes
  B3.2. Figure captions
  B3.3. References

(see List 7 for further treatment)

This text structure information is a clear example of the power of a General Mark-up Language. In inputting the text, all the different identifiable entities are preceded by a tag. Dependent on the use of the stored information, this tag can then be translated into an action (e.g. print quotations in italic) or omitted (e.g. if the quotation is displayed on a normal monitor).

One has to realize that within the normal text elements, codes for special characters and symbols may also appear. These are codes (mostly mnemonics, combinations of normal alphanumerics) indicating most of the time mathematical symbols, greek or other non-latin characters. When reading back the stored information, a translation always has to be made. For typesetting, the coded instruction will be translated, e.g. an α on a normal terminal screen has to be translated into 'alpha'. In this article we deal only with the information elements or SGML structure. It is important to note that, although SGML codes (e.g. the tag 'surname') may involve typesetting instructions (e.g. first character capital and the other characters of the surname lower case), a distinct set of codes for abnormal (e.g. non-ASCII) characters should always be present.

(c) 'Third-order information' concerns the *illustrations* of an article. They are an integral part of the article, but are always treated separately, mainly for technical reasons. Illustrations sometimes give at a glance all the information contained in the article. On the other hand, processing a picture conventionally, or electronically, is of a completely different order of magnitude of effort as compared with text. Therefore this 'third-order information' (Set {C}, List 3) contains all different types of possible illustrations. The captions of illustrations, however, belong to Set {B} as being related text.

List 3. Third-order information
Set {C}. Illustrations
C1. Black and white pictures
  C1.1. Line drawings
  C1.2. Half tones
  C1.3. Photographs
C2. Coloured pictures
  C2.1. Line drawings
  C2.2. Half tones
  C2.3. Photographs/slides
C3. Digitized figures

3.2. Information transfer from author to publisher

All different types of information contained in a journal article can be analyzed this way. Such a formal division does not, however, give us the whole picture. We do not deal with abstract information categories, but with physical manuscripts delivered by an author, so we also have to identify the different ways in which authors submit their work. Generally speaking, we can make a distinction between three ways of manuscript submission:

(a) The classical way of a hand-written or type-written manuscript. There is no intrinsic difference for a publisher in receiving the manuscript produced on a sophisticated typewriter or legibly handwritten, as the information only has to be readable in order to be reprocessed (edited, typeset, etc.).

(b) The next stage in sophistication is an author with a word processor. In this case, there is some preformatting within the text and there are some tags in the word processor file. In very simple situations, the word processor text file can be loaded into the publisher's computer and therefore might reduce the labour involved in processing the article. The complete retyping is not necessary, but all the appropriate editing has to be done. A major obstacle with these ways of submission is that there is no standardization between different systems. The identifiers in a word processor file are not the same as those needed for the typesetting system of the publisher. For example, in a text created by means of a word-processing package, one will often find so-called 'soft carriage return' codes in order to indicate that an 'end of line' is reached, while the normal carriage return is used for indication of an 'end of paragraph'. When this text file is converted to an external form (e.g. on floppy disc) by the word-processing package, all 'soft carriage returns' may be converted to normal carriage returns after all lines have been justified. Extra fixed spaces have to be removed because the typesetting system uses variable spaces, and the 'end of paragraph' cannot be recognized.

So, although a word processor environment offers possibilities, in normal practice with relatively short journal articles, it is easier to reprocess the hard copy instead of using a word processor file.

(c) A third possibility, which we would like to call the ideal possibility, is a situation where the authors use word-processing packages integrated structurally with the requirements of the publisher. This does not mean that the author has to take care of all the tagging and identification. No author will ever be ready to submit his/her manuscript completely according to, e.g. the bulky SGML application of the Association of American Publishers [3], but prestructuring can be easily done. Thus, if a quotation can be identified locally, the author can decide to have it 'bold face' on his private copy for instance, while the publisher, by identifying the tag 'quotation', can print it in italics if that is the house style of the publisher. In the course of this article, we will eventually end up with a list of preconditions for such an ideal system. The different ways of manuscript submission are depicted in Fig. 1.

3.3. *Publishers' information*

After having received the final manuscript in one form or another, many activities are performed by the publisher. The manuscript has to be edited: in many cases this means correcting lan-

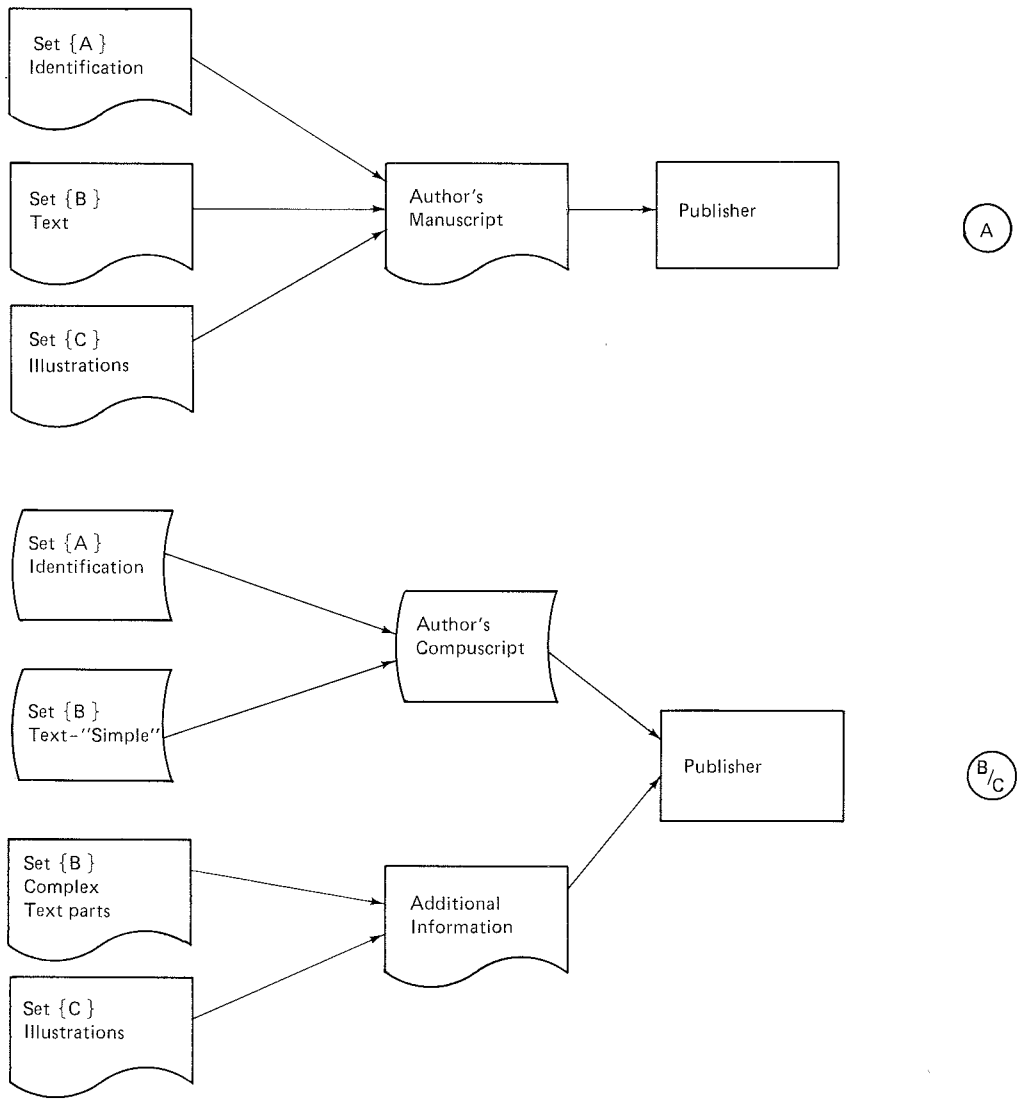

Fig. 1. Various ways of manuscript submission.

guage, systematizing the spelling and structuring the contents in clear paragraphs, sections, chapters, etc. Besides this overall editing, mark-up instruction has to be added. This whole mark-up procedure prepares the manuscript for its next stage in life: the typeset version. Within this activity, the publisher's office adds information to the article, as well as changes existing information, i.e. the information defined in section 3.1 {A, B} so that addresses can be standardized, postal codes can be added, the layout of the text can be changed according to the house style rules, etc. Therefore, although no entities are added or subtracted, the contents may be slightly changed. In this article, we do not consider any procedures relating to changes for illustrations, but, of course, they have to be enlarged, reduced or otherwise made to fit into the final article. The information the publisher adds, given in List 4, contains an article identification number, information on the structure of the article (such as the number of figures, the number of references, the length of the manuscript and the estimated length of the final version) and the document type, e.g. review article, letter or erratum.

List 4. Publisher's information
Set {D}.
D1. Identification number
(such as accession number)
D2. Journal title
D2.1. Abbreviated title
D2.2. ISSN
D2.3. Coden
D3. Size
D3.1. Number of manuscript pages
D3.2. Number of tables
D3.3. Number of figures
D3.4. Number of references
D4. Document type
(review, letter, erratum, etc.)
D5. Dates
D5.1. Date of receipt
D5.2. Scheduled date of publication
D5.3. Actual date of publication
D6. Other production information
D6.1. Editing status (1st proof, 2nd proof, etc.)
D6.2. Production status (for printing, etc.)

A possible treatment of the different types of articles mentioned in section 3.2, is as follows. In the case of handwritten or typewritten manuscripts, the editorial office of the publisher will mark-up the manuscript and subsequently the manuscript will be keyed in by the typesetter. In the case of word-processing output, which is readable on the publisher's machine, the word processor file will be converted as far as possible. The hard copy of the manuscript will be marked-up and the converted text will be edited online. The edited file will then be sent to the typesetter's computers. Online editing is still in its infancy and, at the moment, hardly any publisher or typesetter can deal with a large variety of word-processing packages. A major problem remains the fact that one needs the hard copy of the manuscript for mark-up because the author is free

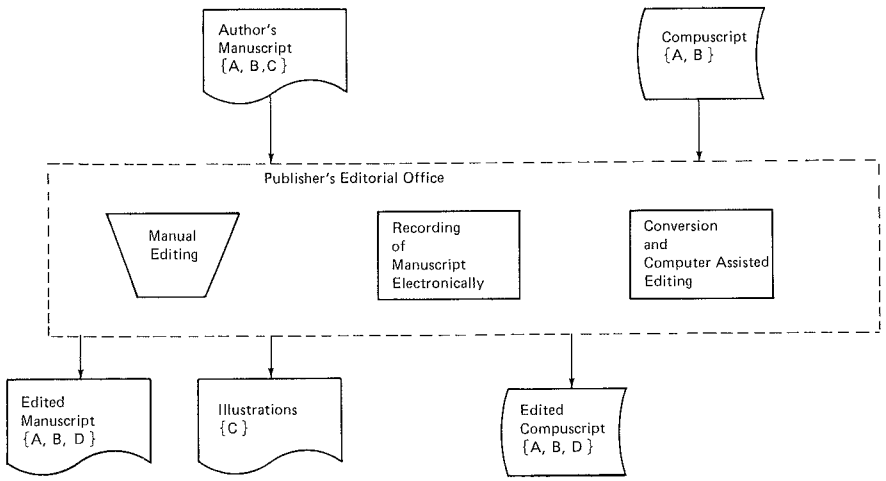

Fig. 2. Manuscript processing at the publisher's office.

Box 2. Proposal for a current awareness service

Between acceptance of a scientific article and the arrival of the journal issue containing this article in the library takes (depending on length and complexity) between 3 and 9 months (for mathematics even much longer). After publication the various abstracting services use the final printed version for their activities. Therefore, another time delay of at least 2–3 months is introduced before people can search online for relevant literature. The only way this serious problem can be solved is that the publisher supplies his readers with information as soon as this information is ready for dissemination. In earlier times, this was only after publishing the article. In our approach, all relevant information for researchers, such as title, abstract, names and affiliations, as well as the final destination, are known and manipulated from the beginning.

So, with a subset of the publisher's database, containing parts of the identification Set {A} and the publisher's information Set {D}, online searchers are able to identify relevant literature. After reading the abstract, one can decide to apply for a preprint (to be ordered online, or from the author) or to wait until the full article is published. The status information of the article will indicate when publication is expected, etc. With a simple query language, researchers will be informed continuously on things like: which articles are in the pipeline for a certain journal, or whether a particular author or group has already published new results, etc.

Within the development of the production, more precise information (e.g. bibliographic data, keywords) can be added, as the current awareness sub-database is continuously updated due to the ongoing publishing activities. This way, for the first time in history, users will be able to be informed of scientific publications from the day of acceptance, rather than from the day of publication. The user of this service has to realize that, as long as the status tag has not yet changed to 'final version', minor mistakes and errors (mis-spellings of names, wrong numbers) may occur. Experience shows, however, that this type of error is minimal for the high quality operations of most large scientific publishing houses.

to do what he/she wishes without any consideration for the publisher's requirements. If, for instance, an author has a lot of quotations and likes to have it bold face italic, the author is perfectly able to do that, while the publisher's house style might be only 'quotes' without any change in font. With all the unclear and non-standardized codes of word-processing packages, online editing will remain a cumbersome activity. The third, ideal situation is, of course, where the word-processing facilities are compatible with the publisher's system, e.g. where an author wants to have a quotation, then the author is free to print it in any style, while the publisher's reads only the identification tag and does not have to worry about the peculiar taste of an author. In this case only, hard copies can be dispensed with. After the key-in of the traditional manuscript or conversion of the electronically submitted manuscripts, we have a situation in which the information given by Sets {A,B,D} is stored in a computer, while the information of Set {C} is somewhere on the shelves in the publisher's office, see Fig. 2. All this stored information is split up into well-defined entities, so, with a proper relational database management system, the publisher is able to create a multiplicity of products.

3.4. Intermediate products

(a) First and foremost, of course, the information will be converted to a phototypesetting or laser printing system to create 'old fashioned' first proofs of the different articles. They will be proof-read by the editorial staff and/or by the authors themselves. In this procedure, not only will mistakes be corrected, but also minor changes will be implemented. Recycling results eventually in the final—completely correct—version. Different elements from Sets {A} and {B} are changed, while from Set {D} the dates and possible other production information are updated.

(b) Secondly, all kinds of secondary information, like author lists, production load, and general statistical information can be generated. This information can be used for, e.g. direct mail, or analyzing the publisher's activities in terms of geographical spread of authors, etc.

(c) Thirdly, at this stage of information processing, we can already develop new products. Owing to the database approach, it is becoming possible now to enable authors and readers to know in advance which articles are due to be published. Part of the information can be regrouped in a sub-database, which can serve as a

publicly searchable current awareness database. The information of Sets {A} and {D} are the most appropriate elements for such a database service. Users can then search for articles to be published in a particular journal, or by a particular author, and in this way can learn about current research months before the actual printed article appears on the library shelves. In Box 2 an idea for such a current awareness service has been worked out. It is important to note the difference between such a service and an online journal. In the case of a current awareness service, only key information is searchable, in the same way as in conventional bibliographic databases. The difference here is that the information is brand new, and that the final bibliographic information is

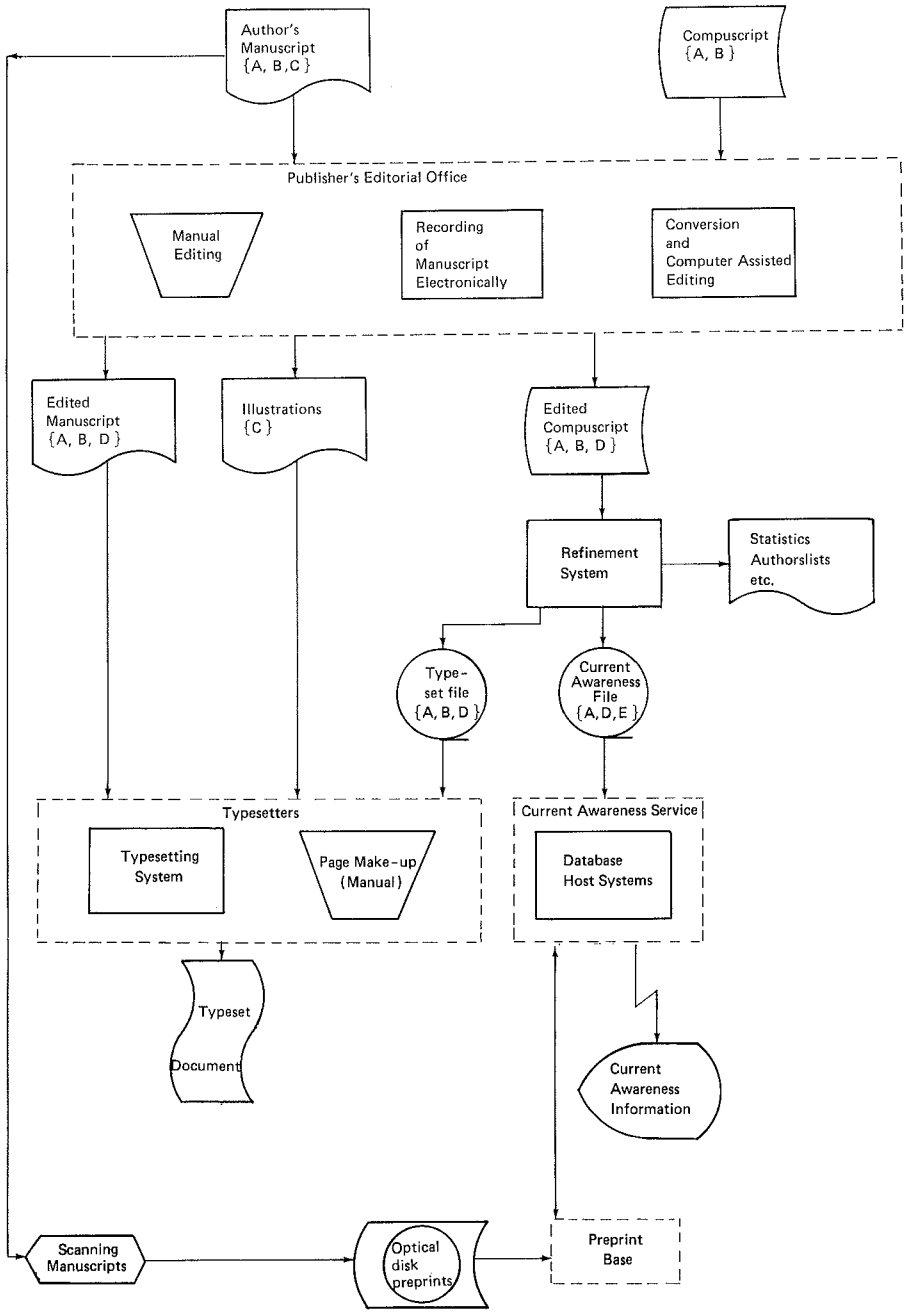

Fig. 3. Flow chart of manuscript handling and intermediate products.

given only in the course of various updates. The first appearance of a reference to an article will indicate only that it is 'in proof' for a certain journal. Only after all proof-reading and correction is the final information available containing all bibliographic references (see below).

Especially in science, where many non-standard alphabets and mathematical expressions, which cannot normally be displayed on a terminal screen, are abundant, only part of the full information can be supplied. This is also the reason why producers of bibliographic databases try to minimize all mathematical formulae or symbols in the abstracts.

In the case of an electronic journal, we deal with the complete text of the article including formulae, tables, etc. Although a current awareness service is a first step towards the possibility of having the complete article online, this step is extremely large and requires very sophisticated text manipulation techniques which are scarcely developed at the moment. (Of course, for very simple flat text, the problem is much easier and will be solved fairly simply in the near future.)

3.5. A preprint base

The creation of another product from the original manuscripts delivered to the publisher can result from digitally scanning the articles entirely (figures included) and storing them in a high density (optical) memory medium. In this case, we simply copy the manuscript onto a memory without any structuring. Such a so-called bitmap representation only enables the user to produce the whole article, which is to be considered as one bitmapped picture, as distinct from the previous approach of a character by character input. Such an (optical) medium is, therefore, not randomly searchable; but every article can be found by its identification number, which is unique. A link between this accession number and the (same) accession number in the current awareness subdatabase can create a completely new form of information dissemination. An end-user (reader and/or author) can search for a particular document in the current awareness database. If, after reading the abstract he/she decides that the information is so important that he/she cannot wait for the final printed version, the manuscript can be requested and, from the (optical) disc, a copy can be retrieved. This copy can then be mailed, or sent by facsimile, to the end-user. The great advantage of such an arrangement is that (optical) discs containing the manuscripts received (say) every month, can be distributed to various main libraries, while the information in the current awareness service will be updated continuously. For instance, if a reader learns from the current awareness database that the article has just been received, he/she can decide to request a facsimile of the manuscript. If the current awareness database indicates that the article is on the verge of being published, he/she can decide to wait for the printed version (which is edited properly and is much more readable than the original manuscript), and, if wanted, order a reprint.

For some people, this option might look a little artificial, but, in some fields of science, sending around preprints is part of the publication culture. In high energy physics, copies of a submitted manuscript are mailed by the author to a few hundred colleagues. Rough calculations show that the total number of preprints sent around in high energy physics amounts to about two million copies per year. A hybrid system of a current awareness service online and an (optical) storage of the complete manuscripts would be of great use to this particular community.

Nobody is bothered with too many preprints of which only part is of his/her interest. On the other hand, an author is sure that his/her colleagues are sufficiently alerted that an article has been written and can be read quickly. An overview of the intermediate products is given in Fig. 3.

3.6. Indexing information

In the course of the whole process, indexing can be carried out. At the moment, articles are indexed by database services, like Chemical Abstracts or Physics Abstracts, only after (primary) publication. Because the publisher's database always contains the latest bibliographical information, one can create a dynamic system in which the classification codes, keywords and other indexing tools are added during the publication process. During the time authors proof-read their articles, indexing can also be performed. The information for retrieval, Set {E}, will then be added into the publisher's database as well as to the sub-set of

the (searchable) current awareness database (see List 5).

List 5. Retrieval tools
Set {E}. Retrieval keys
E1. Controlled language terms
(belonging to a language thesaurus)
E2. Free text terms
(not belonging to a thesaurus)
E3. Classification codes
(e.g. PACS)
E4. Registry numbers
(e.g. chemicals, biological organisms, etc.)
E5. Molecular formulae
E5.1. Linear notation
(e.g. Wiswesser notation)
E5.2. IUPAC
E5.3. Structural formulae

3.7. Bibliographic information

After a number of correction stages the article is ready for publishing. The publisher's database now contains the information from the last updated versions of Sets {A,B,D,E}, while part of this database, which is accessible to the outside world, contains the current awareness information of Sets {A,D,E}, and the complete manuscripts are available on, e.g. an optical disc. At a slightly earlier stage, it has been decided in which volume and in which issue of the particular journal the article will be published. In a new information group (Set {F}, the 'bibliographic information', see List 6), the scheduled volume and issue and publication date can be given, and this information can also be part of the current awareness database.

List 6. Bibliographic information
Set {F}. Final bibliographic references
F1. Volume number(s)
F2. Issue number(s)
F3. Page number(s)
F4. Journal cover date

At the very last stage, Set {F} is completed with the complete bibliographic information, including cover date, etc. At this point, the final articles will be packaged together into a journal issue, printed and mailed to the subscribers. In the current awareness database, these articles will have a tag reading "published in journal X, page Y".

4. Final products

The main final product is, of course, the traditional printed version of the article in the journal issue. The publisher's database, however, contains all information stored in this article except for the figures. (One can think of digitizing the figures and storing them in the database, but this is extremely memory consuming and does not give many extra possibilities for use.) This full text database can create several different types of information:

(1) Publisher-generated information, such as author and subject indexes, contents lists, master indexes, can be created automatically from the database and can be used for publishing in print or online.

(2) The bibliographic database is the mature current awareness database. While in the latter the information is updated continuously until the final stage is reached, the bibliographic database is a cumulating product where, in time, all bibliographic information of all articles ever published will be stored.

The bibliographic database created as a spin-off of the publishing activities can serve as the basis for all kinds of specialist databases. Various institutions which create bibliographic databases for a limited public can make use of the full publisher's base, and have only to add the extra intellectual effort of dedicated indexing.

(3) When we are dealing with flat text without tables, chemical/mathematical formulae or strange symbols, a full text database of articles can be created, which can be uploaded into a special system where full text searches are possible. For much scientific information, this will hardly be possible due to the large amount of non-ASCII symbols and alphabets used. So, depending on the nature of a journal, one can decide to have it available publicly as full text or only the bibliographical parts of it (including abstracts).

(4) A citation index can be made without any problem if the references (see List 7) are already split into their distinct entities.

List 7. References
Subset {B3.3}. Citation information
B3.3.1. Author's name
B3.3.1.1. Surname
B3.3.1.2. First name(s) or initials
B3.3.1.3. Pre or post particles

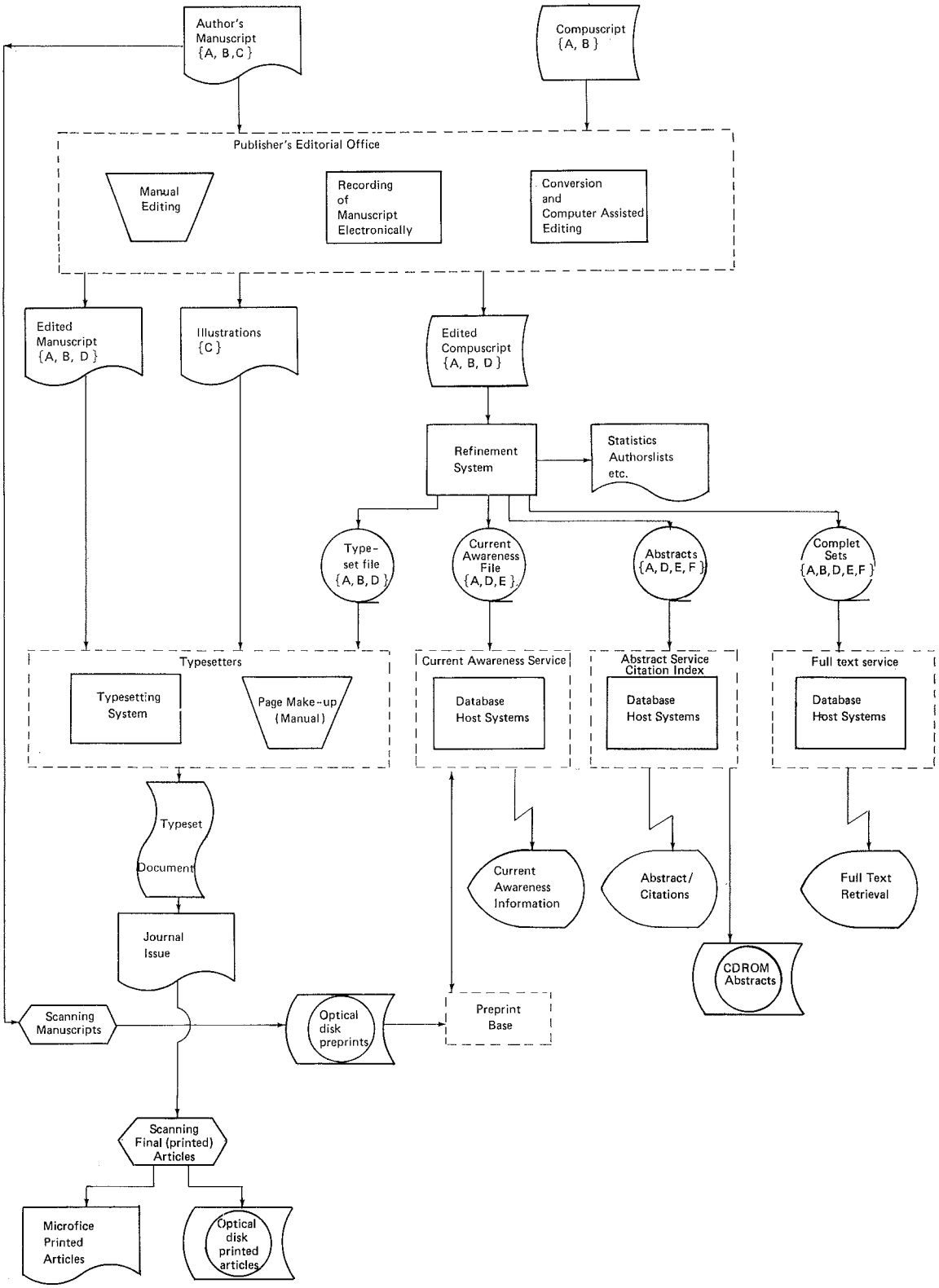

Fig. 4. Total overview of document handling, intermediate and final products.

B3.3.2. Bibliographic information
B3.3.2.1. Journal name or abbreviation
B3.3.2.2. Volume
B3.3.2.3. Year
B3.3.2.4. Page numbers
B3.3.2.5. Book title
B3.3.2.5.1. Chapter title
B3.3.2.6. Publisher's name
B3.3.2.7. City (and country) of publication

(5) The final version of the articles can again be scanned completely and stored onto an (optical) memory medium, thereby creating a complete record of all published material in a very comprehensive and economical way. Again, the connection between the bitmapped version and the character-coded bibliographic database makes quick searching and easy retrieval possible. Libraries might have the option to have regular optical discs or CD-ROMs from the publisher, instead of the printed journals. Then library users can search online for authors, keywords, journal names, etc., and print locally their article on-demand from the electronically stored journals in their own library.

(6) Old-fashioned storage on microfiche can obviously also serve as an additional dissemination possibility.

(7) Parts of the final information, e.g. the abstracts, can also be delivered on other media, like CD-ROMs. A complete picture of the overall structure of this set-up is given in Fig. 4.

5. Recommendations for an author's workbench

The system described above deals only with the publisher's office and its interfaces to secondary services and hosts. A major problem remains, however. As already discussed in section 3.2, the way authors submit their manuscripts is still completely non-standardized. In order to integrate the author's input with the publisher's system, a strict, stratified author's workbench has to be developed. At the moment, various types of word-processing packages are on the market. Some of them are easy to use, but most of them lack a unified structure. Firstly, all packages using a bitmapped representation are useless, because their files cannot be transferred to an organized database. Only those systems in which the construction of text, including formulae, etc., is stratified and has a linear-coded representation make conversion possible. At the moment, besides professional typesetting programs [4] only two user packages are frequently used: TROFF/EQN [5,6] and TEX [7,8].

These programs are a mixture of an abstract typesetting program, where there is no need for a clear representation (the typists are not interested in the form of the formulae) and a user-oriented system, where the author can easily read the input, e.g. typing an infinity sign in a production environment one wants the least keystrokes necessary. In TROFF/EQN one writes "infinity" while in TEX "/infty". Although the author can read this, it is still extremely cumbersome to correct a manuscript and inexperienced or irregular users are confronted with extreme difficulty in its use.

In an author's environment it is, in our opinion, not important that the text-processing package is able to mimic a typesetter's environment, including all typographical pretensions. The aim of the author is to be able to type his/her manuscript as easily as possible, with a direct readable display on the screen, and in such a structured way that no extra difficulties or layout requirements are needed.

First and foremost, an author's system has also to be built around a strictly defined SGML set of entities. Such a program should preferably be menu-driven, so that an author can type, e.g. only his surname in the desired field, etc. Given such a menu-driven input, the layout can be defined afterwards. As is explained above, using SGML, the typographical form can be defined after data capture. So, with a menu-driven SGML-based package, the author can make his/her preferred layout locally, and is not bothered by the requirements of the publisher, nor is the publisher bothered by the aesthetic taste of the author. This point is a major departure from existing packages.

For formulae and other abstract, but structured information, a system-driven design is required. A promising development here is the GRIF project of the University of Grenoble [9–11]. In this package, the cursor movements are defined by the mathematical symbols used, and there is a strict division between the linear way of storing the information, which makes it perfectly transportable, and the representation on normal or graphic screens, where the formulae are displayed truly as formulae. So, besides the requirement of a struc-

tured system which forces the author to keep within this desired framework, the author may only want to see a completely readable version on his/her screen, which can be edited, changed, etc., directly without any need to remember all kinds of sophisticated coding. Of course, a spelling checker and other kinds of author help functions, such as lexicons, buzzword counters, sentence length counters, etc., may be included.

To summarize, the requirements for an author's system are:

(a) self-explanatory without large manuals and all kinds of coding;
(b) menu-driven according to an SGML structure, so that no mistakes can be made, clear identification of the various fields in, e.g. the bibliographic or reference information, as well as in the building structures, such as mathematical or chemical formulae;
(c) as the author is only interested in a normal readable input, the display on the screen has to be in a readable form, without any information on the coding (for formulae this is a crucial point);
(d) given the defined fields, the author must be able to declare layout requirements for local use, if he/she wants to transport the file to, e.g. a laser printer for distributing drafts: these layout requirements have to be defined in a separate set of instructions;
(e) the storage of the manuscript has to be in a strictly linear fashion, only dealing with normal ASCII characters, so that it can be transmitted in all possible ways to the publisher's office.

These considerations still leave out the problem of figures and tables. As explained already, these entities are still too difficult to cope with. Tables, which have little standardization (form, dimensions and labels of the various rows and columns need special study), are especially difficult. One can even wonder if, ultimately, tables will have to be treated like figures: in most databanks, numbers are hardly used in the form in which they were published, but are converted to standardized dimensions.

6. Conclusions

In this article we have attempted to sketch a method of integrating modern media and traditional publishing. Traditional publishing is a very strong and healthy industry, which has proven its capacities and potential over the centuries. The electronic revolution will open completely new and yet unknown horizons. Our aim is to analyze the scientific publishing endeavour, to enhance the existing activities and to allow for all foreseeable new developments. An integrated approach is only possible by first identifying all the distinct information elements within the problem. This information analysis leads straight to the formalized approach of Generalized Mark-up Language as a system for describing information elements which is logical, as well as flexible and extendible. Moreover, the analysis implies the possibility of total management of all the different information elements: the concept of a relational database as the nucleus of all operations is fundamental. Starting with a relational database, traditional publishing is conceptually a subset of existing possibilities, even though traditional publishing makes up more than 90% of the activities. An advantage of this evolutionary approach is that other types of retrieval and dissemination can be allowed to develop without major disturbance of the existing processes. Also in this field some experiments are underway, but an integrated approach for the whole industry is yet lacking. Our aim in this paper was to attempt to identify and conceptualize the necessary steps in solving part of the problem.

Acknowledgements

The critical discussions with, and help of, our colleagues, especially Barrie Stern and Janette Young, are gratefully acknowledged.

References

[1] J.G. Kircz, Will physics publishing survive the electronic challenge?, in: J.J.J. Miller, ed., *PROTEXT II*, Proceedings of the Second International Conference on Text Processing Systems (Boole Press, Dublin, 1985).
[2] ISO, Information processing—text and office systems—Standard Generalized Markup Language (SGML), Draft International Standard ISO/DIS 8879, 1985.
[3] A.W. Kenneth Metzner, *IEEE Trans. Prof. Comm. 18* (1975) 274.
[4] J.W. Seybold, *The World of Digital Typesetting* (Seybold, Media, PA, 1984) and 1985 supplement.

[5] B.W. Kernighan and M.E. Lesk, UNIX Document Preparation, in: J. Nievergelt, G. Coray, J.-D. Nicoud and A.C. Shaw, eds., *Document Preparation Systems* (North-Holland, Amsterdam, 1982).

[6] B.W. Kernighan, The UNIX document preparation tools —a retrospective, in: J.J.J. Miller, ed., *PROTEXT I*, Proceedings of the First International Conference on Text Processing Systems (Boole Press, Dublin, 1984).

[7] D.E. Knuth, *TEX and Metafont* (Am. Math. Soc./Digital Press, Bedford, MA, 1979).

[8] D.E. Knuth, *The TEX Book* (Addison-Wesley, Reading, MA, 1984).

[9] V. Quint, *Tech. Sci. Informatics 2* (1983) 169.

[10] V. Quint and I. Vatton, Grif: un editeur interactif de documents structurés, Raport de Recherche Tigre No. 27, IMAG, Grenoble, 1985.

[11] V. Quint and I. Vatton, Grif: an interactive system for structured document manipulation, in: J.C. van Vliet, ed., *Text Processing and Document Manipulation* (Cambridge University Press, Cambridge, 1986).

[12] Association of American Publishers, Electronic Manuscript Series 1986, (a) Standard for electronic manuscript preparation and markup, (b) Markup for mathematical formulae (draft), (c) Markup of tabular material (draft).